

Standardizing Format Descriptions

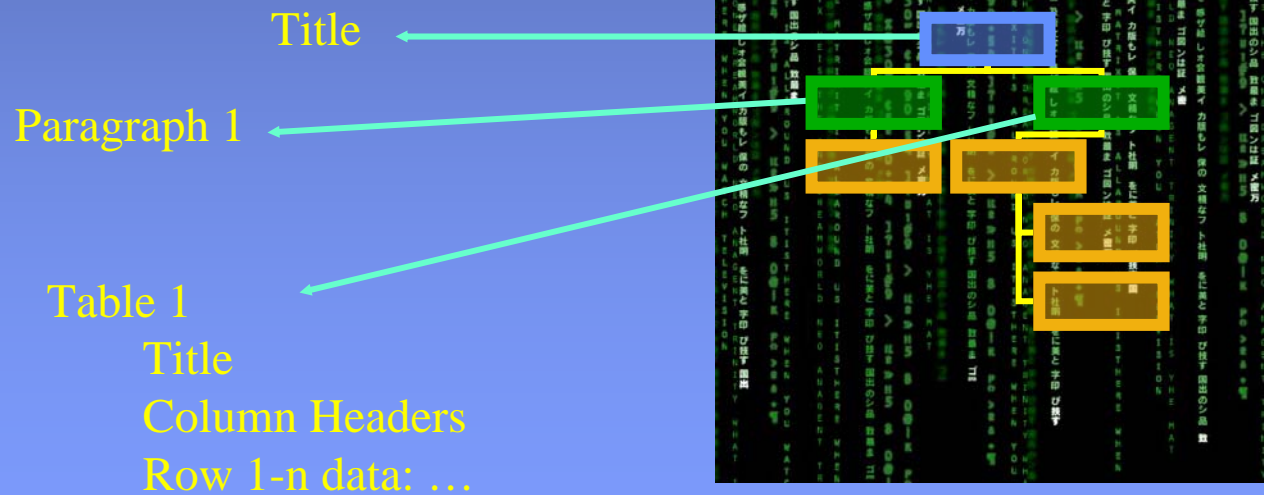
Jim Myers

**Chief Scientist, Computational Sciences
and Mathematics Division**

Pacific Northwest National Laboratory

Concept

- A standardized language for describing in detail the meaningful sub-structures in file formats and associating these structures with bytes in specific files






GGF WG Co-Chairs:
Mike Beckerle,
Ascential Software
Alan Chappell, PNNL
Martin Westhead

Data Format Description Language “Daffodil” Working Group, Global Grid Forum

- Defining an XML-Schema-based language
- Based on numerous existing tools: BFD, BinX, ESML, CML, products from Ascential and IBM
- Supports mapping of arbitrary ASCII/binary file formats to an XML data model

How does DFDL work?

```
<xs:schema elementFormDefault="qualified"
  attributeFormDefault="unqualified"
  xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns="DFDL">
  <xs:element name="myData">
    <xs:complexType>
      <xs:annotation>
        <xs:appinfo>
          <dfdl:byteOrder>bigEndian</dfdl:byteOrder>
        </xs:appinfo>
      </xs:annotation>
    </xs:complexType>
    <xs:sequence>
      <xs:element name="x" type="xs:int"/>
      <xs:element name="y" type="xs:int"/>
      <xs:element name="xdata" type="xs:float" maxOccurs="unbounded">
        <xs:annotation>
          <xs:appinfo>
            <dfdl:runtimeOccurs>../x</dfdl:runtimeOccurs>
          </xs:appinfo>
        </xs:annotation>
      </xs:element>
    </xs:sequence>
  </xs:element>
</xs:schema>
```



```
<myData ...>
  <x>2</x>
  <y>4</y>
  <xdata>2.78</xdata>
  <xdata>3.14</xdata>
</myData>
```

DFDL supports:

- Basic ASCII/Binary Read capabilities
- Reference – use of a previously read value in subsequent expressions
- Pattern recognition – specifying delimiters to recognize fields/structures
- Choice – use of a previously read value to select among format variations
- “Push-back” capability
- Multi-layer – description of an intermediate representation not exposed in the final result
- Multiple input streams
- Inclusion of static info, e.g. “units”
- Defaulted input for missing values
- Basic Math – in DFDL expressions and representations/values
- Input Validation (partly from XML Schema)
- New type/transform specification

Ties to Grids

- **DFDL is a technology that can be embedded in services that translate/transform/subset digital entities on-demand**
 - Translating between communities
 - Performing streaming translation between programs
 - Returning requested structures from within large data sets

DFDL for Preservation

- **DFDL associates ‘opaque’ source data with a well-defined and well-documented content model**
 - Applicable to all ASCII and binary formats
- **Format description is standardized, human readable, and operational:**
 - Generic DFDL parser can be used to dynamically generate tagged version of content for arbitrary formats
 - Transformed content can be validated against model schema
 - XSLT or other tools can convert content for display
 - Generic and format specific